White Paper Report

Report ID: 100938

Application Number: HD5110610

Project Director: Yuri Tsivian (ytsivian@uchicago.edu)

Institution: University of Chicago

Reporting Period: 9/1/2010-3/31/2013

Report Due: 6/30/2013

Date Submitted: 6/28/2013

White Paper

Report #: HD-51106-10

Title of Project: Cinemetrics, a Digital Laboratory for Film Studies

Project Director: Yuri Tsivian

Institution: University of Chicago

Date Submitted: June 28, 2013

WHITE PAPER

Summary

The subject of this report is *Cinemetrics* (http://www.cinemetrics.lv), a website and database designed to supplement the traditional toolkit of film studies. In 2010, Cinemetrics won the NEH Start-Up Grant in Digital Innovation for 18 months. In 2012 the project received a year-long no-budget extension. March 31, 2013 was the end day of the extended grant. The work performed in the 2.5 years of the Grant period we hit most of the targets outlined in the 2010 proposal. On some fronts we did more than we expected, on some less. Database expansion, client tool improvement and statistical analysis are main parts of our success story; we could have been more fortunate with data verification. The principal goal of the Grant has been achieved: the Cinemetrics database, measurement software and analytical capacities have been radically innovated.

Description

1. The principle objective of Cinemetrics as a project is to create a digital platform for informed, comprehensive and collaborative research in the history and theory of film editing. My application proposal for this grant envisaged three major steps towards this goal. Cinemetrics, which was 5 years old at the time I applied, was in need of innovative effort on three fronts: film statistics, client tool improvement and existing data ranking and verification.

The work plan I developed to achieve all this relied on a team of consulting experts, researchers and software specialists. My role was that of a coordinator and, whenever my qualifications allowed, co-researcher and collaborator on par with the members of my team.

I now introduce the team rubricated by their parts in the project:
Gunars Civjans, Cinemetrics website designer and software developer;
Mike Baxter, Nick Redfern, Barry Salt, consultants in statistics;
Arno Bosse, Keith Brisson, collaborators in digital analysis;
Daria Khitrova, collaborator in film-historical analysis;
Ian Jones, research assistant in database cleansing and data verification;
Peter Thorson, code review & optimization and the Chicago crew website maintenance.

2. The second section of the report itemizes and estimates the work performed by the members of the NEH-Cinemetrics team. I start with A) the statistics crew, move on to B) computer scientists, on to C) the research in theory and history of film; I will end this section with D) achievement and problems in the field of database unification and verification.

A) An inherent trait of different cinemetrics analyses is that they all start with quantitative data ending up with qualitative conclusions. This (very general) description highlights the proximity of cinemetrics' mode of operation to statistics. The initial plan of involving expert statisticians was to use them as consultants and supervisors whose role would be limited to answering our (film scholars') statistical questions and catch our possible (indeed, inevitable) statistical errors. As it turned out as soon as I contacted statisticians with our offer, our expectations of their part in Cinemetrics innovation have been too moderate and modest. As it happened without us being fully aware of it, academic statisticians were already watching Cinemetrics. Before I had a chance to officially offer academic statistician Mike Baxter the job of consulting Cinemetrics, he contacted me in order to alert Cinemetrics users to the essay he published in a major journal of statistics *Significance* in which Baxter discusses and evaluates statistical

resources and endeavors in the field of cinemetrics (the non-capitalized version of our site is now used as a generic term for the field Cinemetrics helped launching). Cinemetrics Database, Baxter informs his colleagues, provides a large sample of data that will allow statisticians to ask new questions and test statistical tools, see http://www.significancemagazine.org/details/webexclusive/2784341/Cinemetrics-measuring-movies.html and http://onlinelibrary.wiley.com/doi/10.1111/j.1740-9713.2012.00599.x/abstract

It then occurred to Gunars Civjans and me that perhaps our initial plan was not ambitious enough. If statistics is interested in cinemetrics as much as Cinemetrics is in statistics, why not open up a forum which would involve film scholars into a dialogue with experts in statistics? As a result, a new space on Cinemetrics was created which houses an intense conversation between our team of 3 statisticians and myself as a film historian and moderator. This fourfold dialogue which started in the summer of 2012 and is still going on is something we are quite proud of: it is for the first time that a number of questions about the use of film data for statistics and relevance of statistics to cinema were systematically posed and perused from a number of different perspectives.

Those interested will find the entire conversation here:
http://www.cinemetrics.lv/dev/on_statistics.php ; and here is a brief summary of its points. The conversation between the three members of our statistics team is grouped around 3 questions. Question 1 is which average shot length value, median of mean, is more effective as far as film statistics. The choice is more difficult than it may look at first. Medians are more robust statistically, but these can only be calculated using computerized counting tools like ones offered to users by Cinemetrics. On the other hand, a number of film scholars have been calculated average mean values unarmed with any computers, for the simple arithmetic mean can be obtained by dividing the length of the film in seconds (or feet) by the number of shots in it. Question 2 was about which type of statistical distribution tells us what about styles of editing; Question 3 focuses of various ways of comparing similar films (for instance, films by the same genre or director) across the database so that we can see if the same-group films display similar editing patterns.

3. The third section of this report focuses on the work done on the Cinemetrics as a website, its data-collecting tools and the database it hosts. Operations in this area were performed or supervised by Gunars Civjans who designed and created the Cinemetrics website back in 2005 and continues to maintain and improve various facets of its stratified structure which includes exploratory Labs http://www.cinemetrics.lv/labs.php , the Discussion Board http://www.cinemetrics.lv/discussion.php and the cyberspace for methodology debates called Measurement Theory (see http://www.cinemetrics.lv/articles.php# or http://www.cinemetrics.lv/dev/on_statistics.php).

As envisaged in the Grant proposal, in addition to the website maintenance, Gunars' two major tasks in the field of Cinemetrics' innovation included A) optimizing the codes and solving digital "real estate" problems; B) developing a new client tool for the frame-accurate data input. As was decided, two specialists from the University of Chicago Humanities Computing, Keith Brisson and Peter Thorson, were assigned to work as Gunars' team members. This work was supervised by the then Director of Humanities Computing Arno Bosse.

A) During the first year of the Grant, Gunars, Peter and Keith worked on two mutually unrelated tasks. One was to rewrite the initial Cinemetrics codes so that the way each film's ASLs data were stored did not tie the latter to each specific title. In other words, what we needed to achieve is that Cinemetrics data could be detached from their metadata. The reason why we needed this database restructuring

was to be able to group and compare data across the database. The restructuring was achieved by the end of the first year of the Grant, and this enabled Keith and me launch our first experiments which I described in my "log book" in the comment box in this Lab http://www.cinemetrics.lv/lab.php?ID=119 . A more detailed and analytical description of Keith's and mine experiments is found in my essay "Question 3: Looking for Lookalikes?" in the space for the Film and Statistics discussion: http://www.cinemetrics.lv/dev/on_statistics.php

The other task Gunars achieved helped by Arno Bosse and Peter Thorson was the Cinemetrics website relocation form a commercial server in Latvia (its place of birth) to the Humanities server(s) – a less trivial mission than one may think, for it entailed bits of reprogramming and problem-solving. Why we needed this to be done was the speed, reliability and higher capacity of the University's digital infrastructure. The Cinemetrics.lv is now physically in Chicago and will remain there as long as my employment continues.

B) The original (now dubbed "classic") Cinemetrics measurement tool looks like this:
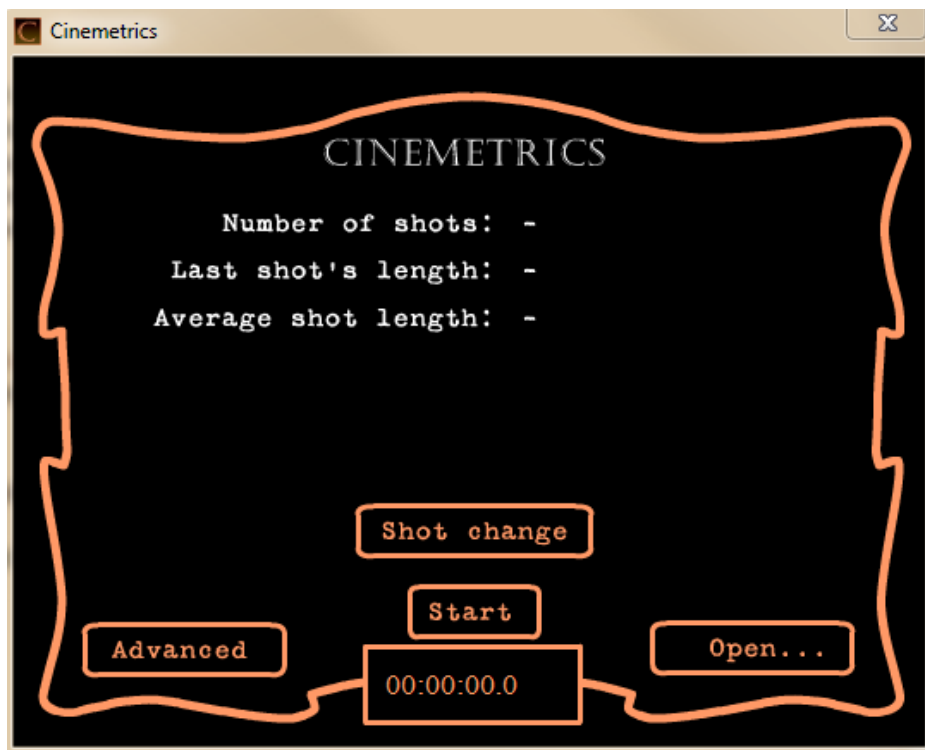
This simple interface works like a stopwatch: it marks time each time you click on your mouse or keyboard button. Normally, one clicks on cuts between shots. The software must run simultaneously with the movie the user is watching. In other words, to measure a film, one watches it clicking on "Shot change" each time one sees a cut. The resulting data are stored in the database and analyzed statistically. The tool is easy to use in any venue, from home to film archives and theaters. The only problem is accuracy. The human reaction time makes it likely that some cuts are missed and some clicked on a little late.

Gunars' task was to design a new-generation measurement tool which we called FACT: Frame Accurate Cinemetrics Tool. The way it works is explained in this Gunars-performed video tutorial
http://www.youtube.com/watch?v=tUAdES1gWgM

Gunars is now in the process of designing a non-video tutorial (not online yet). Here are the basic windows that explain how the FACT works:

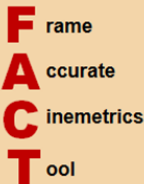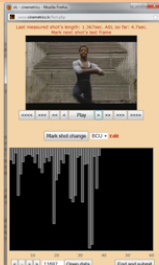### How it works
To use this version of the mesurement tool a copy of the movie you want to measure needs to be saved on the hard drive of your computer. If you only have a DVD of the movie, you will have to rip it first. FACT will open the film inside your browser and provide controls to time to a frame the spot where shots change.

Compatibility
The tool is compatible with most modern web browsers, provided they have the latest VLC media player plugin installed.

How to use it
0. Make sure your browser has the latest VLC media player plugin installed.
1. Watch the video tutorial.
2. Check with the **written tutorial** if lost.
3. Click here to open the tool.

F rame
A ccurate
C inemetrics
T ool

Prerequisites. Steps. Options
Demo on video tutorial

Prerequsites:

1: make a digital video file of the movie you want to work with and save it on your hard drive. **Rip** or convert if your sourse movie is on a DVD. Extensions: .avi bla bla bla will work. Make sure you remember in which directory your video file is saved
2: install the **VLC media player plugin** on your browser.

Steps:

Step 1: **Click here** for the FACT startup window to appear on your screen.
**Step 2:** To load the video file you want to measure, enter the full path and the filename into the textbox. To keep it short, you may choose to save the file in the root directory, for instance, c:/filename.avi
Step 3: "Pause" your movie as soon as it appears on the upper screen; consult the prompts above it for what to do next. Tweak frame controls to find the *first frame* of the first shot. NOTE: you only need to look for the first frame when you start measuring. After that you will be looking for the *last frame* of each next shot to mark.
Step 4: Pause as soon as the shot ends; tweak frame controls < or > to locate and mark the last frame of the shot. As soon as the last frame of the shot has been marked, a shot bar that represents its length appears in the left upper corner of the data graph below. Unlike in your typical histogram, shot bars in Cinemetrics are stalactitally, not stalagmitally attached to the horisontal axis. It is fluctuations in cutting rates, not variations of ASLs that interests us here.
**Step 5:** Click on "End and submit" when the last frame of the last shot has been marked. Fill out metadata and signature boxes on the FACT endup window and "Submit." Go and see the result streight from the thanks window.
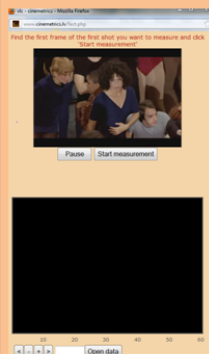
Options:

Option 1: Use the selection box to the right of the "Mark shot change" button to tag shots with the default "Shot scale" or user-defined markers. Do not forget to choose "Advanced mode" on the endup screen if you use this option.

Option 2: You may also use the histogram graph to view, adjust or retag shots. Click on a bar to see the last frame of the corresponding shot appear on the videoscreen. Use controls to zoom in/out of the histogram or to move it to the right or to the left.
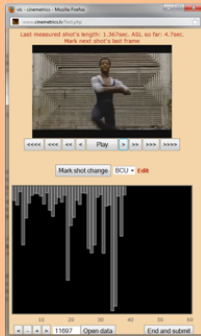
Option 3: Your measurements submitted, the system retains its data as a template. You can reuse your submission at a later time as ofter as you need if, for instance, you want to reeadjust cut times in a different version of the film or retag its shots. To do so, repeat Step 1 using the videofile you once submitted and enter the movie ID found in the URL box of its previous submission. The movie ID are the digits after the "=" sign. Copy those and paste them in the "Open data" box under the hystogram screen. Open the data by clicking the box button and change lenghts or tags clicking on shot bars first. End and submit. The systems saves the new set of data with a new movie ID assigned. Make sure you describe the new type of measurement in the comment box on the endup screen if you use this option.

As follows from the above, the advantage of the new tool over the "classic" one is accuracy; its disadvantage is that it takes longer time to measure (estimated 8 hours per three-hour feature). It is to make the process of measurement faster while keeping the result accurate that Gunars started working on the third-generation tool equipped with automatic shot detection capability as soon as the FACT was finished and made available to Cinemetrics users in spring, 2012. A crucial feature of the tool-in-making will be its ability to measure not only the frequency of cuts, but also the speed of character/objects movements within the shot using software that allows registering and ranking changes between frames.

4. If the performance was affected through Keith Brisson's finding a different job and quitting the University of Chicago it was in the sphere of data ranking and verification. Let me briefly explain what this means. The way Cinemetrics has been designed from the outset was to provide tools and storage for film researchers interested in quantifying the material they worked on. This, in fact, meant (and still means, since we have no plans to change the philosophy of Cinemetrics) that Cinemetrics is a self-service site, a site which relies on the intellectual integrity of those who choose, measure and submit their data to our common pool. You submit your data, you work with your data as your own project requires, but your data are also available to others. It is when it comes to others that the problem of verifying data arises. I can trust my own submissions, or at least I know their margins of error. But to what extent I can rely on your data depends on many factors: your reputation in the scholarly community, your experience with the cinemetrics tools, the accuracy of your submissions overall, etc.

It is our performance on that front that this (fourth) section of the present report accounts for. Early on in the project Keith Brisson grew keenly interested in finding an optimal solution as to which film data on the Cinemetrics database are full and reliable and which are less reliable and partial. Brisson has outlined his plan of solving this issue as a precise and itemized plan he posted on our Discussion Board on 2011-05-04: http://www.cinemetrics.lv/topic.php?topic_ID=360 According to the proposal, Keith was supposed to create a "profiling" piece of software that would rank data automatically using a

number of criteria: a) the length of the submission as compared to the length of this films as marked on the IMDb; b) the number of films earlier submitted by the person who had submitted this particular film; c) how this particular submission tallies with other submission of this particular film.

In order to bridge the gap left by Brisson's departure, Ian Jones was hired as a research assistant responsible for inspecting and cleaning the database. The task assigned to Ian was four-fold: deletion, segregation, ranking and unification. A Cinema and Media graduate student at UofC, Jan is not a programmer, so he was not able to take over Keith's assignment of creating a piece of software to do parts of this job automatically. As we agreed, Ian would do the preliminary "manual" work in this field:

Action 1: DELETION
Identify and delete from the database:
   a) "test submissions" marked so (search for "test" and other possible synonyms; copy such movie's IDs (e.g.=13206) and send to Gunars for deletion;
   b) do the same for submissions with 0 values in ASL or MSL (search by ASL) and send IDs to Gunars;
   c) Use the search method to identify movies without titles;
   d) Identify movies without submitter or year; assess the validity of their data.

Action 2: SEGREGATION
Identify:
   a) films which are not films (for instance, OBAMA INFOMERCIAL (2008), BASEBALL WORLD SERIES, GAME 3 (2006, USA)) or PRESIDENTIAL STATE OF THE UNION ADDRESS (MSNBC) (2007) or AMERICA'S NEXT TOP MODEL series of shows submitted by Christina Petersen; also, search by suspect words like "show" etc.
   b) films measured not by cuts by other criteria; e.g. Tsivian-submitted Chaplin films measuring the laugh frequency in the auditorium, like ID=613, in 2007;

Action 3: RANKING BY ACCURACY
Identify:
   a) Frame accurate submissions by: Barry Salt, James Cutting, Adriano Apra & Simone Starace together (not Starace separately); Heidi Heftberger; Films: TURKSIB ID=8283 and ID=8284; THE HOUSE OF HATE (EPISODE 1); also: Eric T Jones and Viktorija Eksna reported some frame accurate submissions, please check in comment boxes; Also, all films submitted by FACT (check them one by one, there can be bugs and failures)
   b) Mark full film data as separate from fragments data
   c) Rank submissions by reliability:
      Frequent clients vs. one-time clients;
      Researchers vs. Students (Holland, Taiwan, Czech republic; usually the whole class submits the same film)
      Submissions with shots of minimal length 0.1 & 0.2 are suspect of double-clicks ;

Action 4: UNIFY METADATA
   a) USA – United States
   b) Russia before Oct 1917 Soviet Union 1917-1991; Russian Federation 1991-till now
   c) Add missing metadata.

Hired to replace Brisson late in the term of the Grant, Ian Jones was able to do perform Actions 1 and 2.

9

5. This section of the performance report recounts the Cinemetrics activities on the film history front. To work on this front, I enlisted three collaborators: computer scientist Keith Brisson, statistician Mike Baxter and performance scholar Daria Khitrova. My research and experimentation with regards to D.W. Griffith's editing style in collaboration with Keith Brisson and (later) with Mike Baxter has been mentioned in sections 2A and 3A of this report; more details are available on four Cinemetrics Labs (look for keyword "Griffith" in Lab names) and in the "On Statistics" space found under "Measurement Theory" (look on my "Question 3: Looking for Lookalikes?").   Keith Brisson, Arno Bosse and me (in absentia) presented preliminary results in a paper on Cinemetrics given at the Digital Humanities Conference at Stanford in June 2011.

My work with Daria Khitrova was in comparative editing studies. Daria and I explored three massive arrays of data: A) all Charlie Chaplin films made between 1914-1917 at Keystone, Essanay and Mutual studios; B) nearly all available Keystone films released between 1913 and 1916; C) Biograph films directed by Mack Sennett. The number of films submitted by Daria to Cinemetrics equals 192, most of these in advanced modes (exterior vs. interior shot lengths; shot lengths depending on the number of characters involved); some of them were submitted by FACT.

Daria's work was not confined to measuring films and analyzing the results. She did a great deal of field research on the Mack Sennett papers at the Academy Library in LA; watching rare Keystone prints at the UCLA Film Archive in LA and National Film Archive in Prague.   The results of our joint research were presented as a paper "Cinemetrics Looks at Acting: Cross-cutting and Cross-action" at the Berlin conference of the Society for Cognitive Studies of the Moving Image this June, see http://gwk.udk-berlin.de/scsmi/programme_friday.html

6. This section is dedicated to dissemination and impact of Cinemetrics: A) audiences, B) evaluation, and C) continuation of the project.

A) Cinemetrics audiences can be defined in two different ways: *clients* who submit data or texts and nameless *users* who visit the site to look up things. The latter group can be counted by the average numbers of views per day. Their growth between 2010 and now looks like this:

USERS' visits to Cinemetrics have almost doubled during the Grant period:
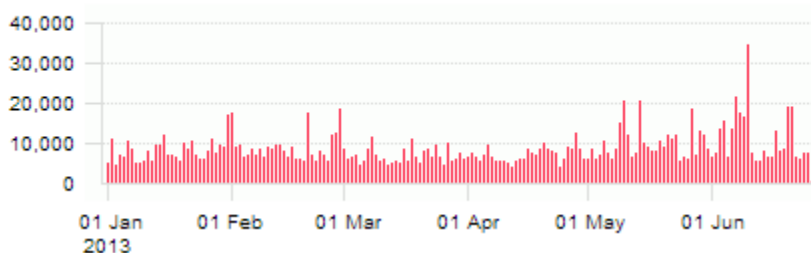July 2010: average page views per day = 4500
Jan 1 – Jun 25, 2013: average page views per day = 8326

And here is the dynamics of page views with the last 177 days (Jan 1 – Jun 25, 2013):
Total page views = 1,473,872

CLIENTS. The cumulative number of Cinemetrics clients increases every day. As I am writing this (June, 28, 2013) our database counts 1,742 names under "submitted by." It would take separate calculation to assess the flowing average growth: some come and go, others come and become our regular clients. But the increase of submitted films tells us indirectly how fast the growth of clients is. This is what I wrote In my 2010 proposal: "Presently the database counts more than 5,000 titles submitted by hundreds of film researchers from different countries." This meant that our average submission rate was 1,000 titles per year starting from 2005 when Cinemetrics was launched. Today the database consists of 12,486 films. This means that in the Grant period (2.5 years) the average number has increased to 3,000 submissions per year, that is, tripled.

The client pool demographics are hard to profile. One thing we know that there are film theory classes in Holland, Czech Republic and (recently) at the National Cheng Ku University in Taiwan that are assigned by their teachers to use Cinemetrics as part of learning about editing. As to the geography within the Cinemetrics reach, the correct answer will be: everywhere on www.

B) Here is a list of interviews and review which shows the spread of Cinemetrics evaluations within the space of the Grant years. There are links I provide for you to check, but my summary is: all reviews are positive.

Reviews and responses. On the list below, items 1 and 3 appeared in film-related press, 2 in the UofC alumni magazine, 4 in the general cultural online journal, 5 in a special journal on statistics:
1. Evan Davis, "CineMetrics," in *The Slant Magazine*
http://www.slantmagazine.com/house/2010/07/its-alive-the-top-film-criticism-sites-an-annotated-blog-roll-part-one/
2. Lydialyle Gibson, "Dancing with Films," *The University of Chicago Magazine,* January-February 2012
http://mag.uchicago.edu/arts-humanities/dancing-films
3. Jesse P. Finnegan, "Measure for Measure," in *Filmcomment,* July-August 2012
http://www.filmcomment.com/article/site-specifics-cinemetrics.lv
4. Vasilii Koretskii, "Genotype of a Film [Genotip filma]" in Russian in: *Colta.ru* Aug 2, 2012
www.colta.ru/docs/3399
5. Michael Baxter, "Cinemetrics: Measuring Movies," in *Significance,* Oct 24, 2012
www.significancemagazine.org/details/webexclusive/2784341/Cinemetrics-measuring-movies.html

Interviews. Of the two different interviews with Yuri Tsivian and Daria Khitrova, one is in Russian, the other in English:
1. "What Is Montage? [Chto takoe montazh?]" in *Colta.ru* Aug 2, 2012 http://www.colta.ru/docs/4210
2. "On Cinemetrics, Video Essays, and Digital Scholarship – An Interview with Dr. Yuri Tsivian and Dr. Daria Khitrova," in *Mediascape* Winter 2013.
www.tft.ucla.edu/mediascape/Winter2013_Cinemetrics.html

C) In March 2013 I applied for a grant offered by the Neubauer Collegium Inaugural Research Project
http://news.uchicago.edu/article/2013/03/04/neubauer-collegium-selects-inaugural-research-projects
In May I learned the good news that my proposal "Cinemetrics Across Boundaries: A Collaborative Study of Montage" has won a two-year award that will enable me to bring to the campus as Visiting Fellows two of my collaborators earlier employed through the NEH Start-Up Grant: Mike Baxter and Daria Khitrova for research and working on the book on the history of editing. We will also use the Collegium funds to call the first Cinemetrics Conference. See more:
https://neubauercollegium.uchicago.edu/page/visiting-fellows